



Sistema de Recomendação para clientes de vídeo locadoras baseado em redes SOM

Trabalho de Conclusão de Curso
Engenharia da Computação

Anderson Berg dos Santos Dantas

Orientador: Prof. Fernando Buarque de Lima Neto, PhD



Anderson Berg dos Santos Dantas

*Sistema de Recomendação para clientes de
vídeo locadoras baseado em redes SOM*

Monografia apresentada como requisito parcial para obtenção do diploma de Bacharel em Engenharia da Computação pela Escola Politécnica de Pernambuco - Universidade de Pernambuco.

Orientador:

Prof. Fernando Buarque de Lima Neto, PhD

DEPARTAMENTO DE SISTEMAS E COMPUTAÇÃO
ESCOLA POLITÉCNICA DE PERNAMBUCO
UNIVERSIDADE DE PERNAMBUCO

Recife - PE, Brasil

Novembro de 2009

Resumo

Sistemas de recomendação têm sido largamente utilizados por sítios de comércio eletrônico como forma de cativar clientes facilitando o processo de compra e resolvendo o problema da sobrecarga de informação. Este trabalho desenvolveu uma abordagem de sistemas de recomendação para o ambiente de vídeo locadoras, onde a busca por recomendações é frequente, mas as opiniões nem sempre agradam o gosto do cliente. Para produzir boas recomendações é necessário criar um perfil para cada cliente. Numa vídeo locadora as informações disponíveis sobre um cliente para construção de um perfil estão em seu histórico de locações. No presente trabalho foi desenvolvido um sistema de recomendação baseado no conteúdo dos filmes presentes no histórico de locação de clientes de vídeo locadoras. Mapas auto-organizáveis de Kohonen foram utilizados para realizar o aprendizado e consequente criação dos perfis de usuários. As redes SOM (Self-Organizing map) organizam topologicamente os filmes, agrupando-os conforme semelhanças entre si. O objetivo é auxiliar o cliente no processo de locação, direcionando-o a realizar uma boa escolha. O sistema obteve bons resultados, deixando o cliente satisfeito na maioria das decisões tomadas.

Abstract

E-commerce sites use recommender systems as a tool for making recommendations to customers. Recommendations make the buying process easy and fun, besides it can resolve the information overload problem. This work developed a recommender system approach for dvd rental stores. Make good recommendations for customers of this type of store is a hard task. Usually, the recommendations made by "word of mouth" are not so good for the customer. It's necessary to know more about this customer to help him make a good decision. Our approach build a user-profile from the rental history of the client in the store. Self-organizing maps were used to build the user-profile. The SOM algorithm organize movies topologically, clustering them by similarities. The central goal is help the user to make a good choice. The system performed well, satisfying the client in the most tests.

Dedicatória

A Deus e minha família, pois me ensinaram os passos que devo seguir.

Agradecimentos

Agradeço a Deus pelo amor e ajuda a todo momento durante a graduação.

Aos meus pais que sempre me apoiaram e me encorajaram a cursar uma faculdade.
Agradecimento especial à minha mãe, que tem sido uma forte coluna.

Ao meu irmão pela compreensão e ajuda e à minha irmã pelo afeto.

Agradeço aos colegas e professores pela confiança e credibilidade que me ajudaram a prosseguir.

Ao meu orientador que acreditou a todo momento que era possível realizar este trabalho.

Sumário

Lista de Figuras

Lista de Tabelas

1	Introdução	p. 11
1.1	Caracterização do Problema	p. 11
1.2	Motivações	p. 12
1.3	Objetivos e Metas	p. 12
1.4	Organização do Documento	p. 12
1.4.1	Capítulo 2: Revisão Bibliográfica	p. 12
1.4.2	Capítulo 3: Modelo Proposto	p. 13
1.4.3	Capítulo 4: Configurações dos Experimentos e Análise dos Resultados	p. 13
1.4.4	Capítulo 5: Conclusão e Trabalhos Futuros	p. 13
2	Fundamentação Teórica	p. 14
2.1	Sistemas de Recomendação	p. 14
2.1.1	Técnicas de recomendação	p. 16
2.1.1.1	Filtragem baseada em conteúdo	p. 17
2.1.1.2	Filtragem colaborativa	p. 19
2.1.1.3	Filtragem híbrida	p. 20
2.1.2	Confiança em sistemas de recomendação	p. 21
2.2	Mapas auto-organizáveis	p. 22

2.2.1	Visão Geral	p. 22
2.2.2	Treinamento	p. 24
2.2.3	Aplicações	p. 26
2.2.4	Trabalhos relacionados	p. 28
3	Modelo proposto	p. 29
3.1	Visão Geral	p. 29
3.2	Implementação do modelo	p. 31
3.3	Funcionamento do modelo	p. 33
4	Experimentos e Análise de Resultados	p. 36
4.1	Base de dados	p. 36
4.2	Resultados	p. 37
5	Conclusões e Trabalhos Futuros	p. 40
5.1	Conclusões	p. 40
5.2	Dificuldades e trabalhos futuros	p. 40
	Referências	p. 42
	Apêndice A: Tabelas de resultados	p. 44

Lista de Figuras

1	Avaliações de usuários no sítio da Amazon.com para determinado produto.	p. 16
2	Recomendações da Amazon.com de acordo com o histórico do cliente.	p. 17
3	Associação de produtos por clientes na Amazon.com	p. 17
4	Córtex cerebral humano	p. 23
5	Rede SOM bidimensional	p. 23
6	Função chapéu mexicano	p. 24
7	Exemplo de vizinhança, onde o instante 2 é menor que o instante 1, que por sua vez é menor que o instante 0	p. 26
8	Resumo do algoritmo de aprendizado SOM	p. 27
9	Fluxograma do algoritmo de aprendizado das redes SOM	p. 27
10	Fluxo de execução do modelo proposto	p. 31
11	Diagrama de classes do modelo	p. 32
12	Diagrama de seqüência da fase de treinamento	p. 32
13	Mapa de filmes do cliente 1	p. 35
14	Mapa de filmes do cliente 2	p. 35

Lista de Tabelas

1	Vantagens e desvantagens das filtragens baseada em conteúdo e colaborativa	p. 20
2	Resultados para o cliente 1 e o filme <i>Shangai Triad</i>	p. 38
3	Resultados para o cliente 1 e o filme <i>The Usual Suspects</i>	p. 38
4	Resultados para o cliente 2 e o filme <i>Mighty Aphrodite</i>	p. 38
5	Resultados para o cliente 2 e o filme <i>Apt Pupil</i>	p. 39
6	Resultados para o cliente 6 e o filme <i>Il Postino</i>	p. 39
7	Resultados para o cliente 6 e o filme <i>Pulp Fiction</i>	p. 39
8	Resultados para o cliente 1 e o filme <i>Richard III</i>	p. 44
9	Resultados para o cliente 1 e o filme <i>Postino, Il</i>	p. 44
10	Resultados para o cliente 1 e o filme <i>Rock, The</i>	p. 44
11	Resultados para o cliente 1 e o filme <i>Platoon</i>	p. 45
12	Resultados para o cliente 1 e o filme <i>Platoon</i>	p. 45
13	Resultados para o cliente 2 e o filme <i>Antonia's Line</i>	p. 45
14	Resultados para o cliente 2 e o filme <i>The River Wild</i>	p. 45
15	Resultados para o cliente 6 e o filme <i>Apollo 13</i>	p. 45
16	Resultados para o cliente 6 e o filme <i>Forrest Gump</i>	p. 46
17	Resultados para o cliente 6 e o filme <i>The Remains of the Day</i>	p. 46
18	Resultados para o cliente 6 e o filme <i>Aladdin</i>	p. 46
19	Resultados para o cliente 11 e o filme <i>Dead Man Walking</i>	p. 46
20	Resultados para o cliente 11 e o filme <i>Braveheart</i>	p. 47
21	Resultados para o cliente 11 e o filme <i>Bob Roberts</i>	p. 47

22	Resultados para o cliente 11 e o filme <i>2001: A Space Odyssey</i>	p.47
----	---	------

1 Introdução

1.1 Caracterização do Problema

A tecnologia, principalmente a internet, tem mudado a forma de fazer negócios na indústria do entretenimento. Vê-se atualmente, em grande volume, a migração do mercado físico (tradicional) para o virtual. Muitas lojas disponibilizam os seus produtos à venda através de sítios na grande rede de computadores, outras, mais ousadas, possuem apenas lojas virtuais. Uma das vantagens dessas lojas é o fato de, por não precisar de um ambiente físico para vendas, o número de produtos oferecidos é muito maior (1), oferecendo assim uma maior variedade de produtos a seus clientes. Porém esse grande volume de produtos leva a um problema conhecido como sobrecarga de informação.

Diante da diversidade de produtos, o cliente que quer realizar uma compra em sítios de comércio eletrônico, freqüentemente precisa de auxílio para encontrar o que deseja. Além das ferramentas de busca, as grandes lojas virtuais disponibilizam uma forma de mostrar ao cliente informações personalizadas sobre produtos que podem interessá-lo, que é o sistema de recomendação. Os sistemas de recomendação podem sugerir produtos utilizando diversos aspectos, como compras anteriores de determinado cliente ou opiniões de outros clientes sobre os produtos da loja. Desta feita, os sistemas de recomendação criam lojas personalizadas para o perfil de cada cliente. A personalização fomenta algo muito importante para o negócio que é a fidelização do cliente (2).

Em vídeo locadoras, por exemplo, é comum a dificuldade de sugerir novos filmes para clientes, mesmo para os mais antigos e/ou mais assíduos. Isso por não ser fácil para o funcionário de um estabelecimento desse segmento identificar o perfil do cliente a partir de filmes já locados e conseqüentemente quais são suas preferências. O cliente, por muitas vezes, segue a opinião de outras pessoas que indicam um determinado título, afirmando ser aquele um bom filme. Opinião que, geralmente, pode não corresponder ao seu perfil. Atualmente as locadoras de DVD estão disponibilizando locações através de páginas na internet com a vantagem da entrega em domicílio. Fato que dificulta ainda mais a obtenção

de opiniões de terceiros pelo cliente. Sistemas computacionais de recomendação podem trazer todas as suas vantagens para vídeo locadoras tanto físicas como virtuais, auxiliando o cliente a fazer a melhor escolha. Isso leva a um fator diferencial da vídeo locadora na visão do cliente.

1.2 Motivações

O que motivou o presente trabalho foi a possibilidade de tornar a escolha de um filme para locação uma experiência mais simples e interessante. O cliente poderá direcionar suas escolhas a partir das informações que o sistema irá fornecer, todas baseadas no histórico de títulos locados na loja. Este trabalho tenta minimizar o problema da sobrecarga de informação provendo o cliente com parâmetros que possam identificar o seu perfil e balizar melhor sua decisão.

1.3 Objetivos e Metas

O objetivo deste trabalho é desenvolver um sistema de recomendação personalizado para clientes que freqüentam vídeo locadoras. O sistema visa facilitar e tornar mais apropriada a escolha de um filme pelo cliente, diminuindo as chances de desperdício de seu dinheiro e tempo em algo que não lhe agrada. Além de oferecer um serviço diferenciado ao cliente, a vídeo locadora irá se beneficiar, também, por uma muito provável maior fidelização do mesmo, já que ocorre uma personalização do serviço. Temos como meta construir a aplicação baseada em redes neurais artificiais, especificamente mapas auto-organizáveis de Kohonen. Essa arquitetura permite que haja um agrupamento do histórico de locações de um determinado cliente de maneira que este possa ter informações visuais sobre seu histórico. Quando o cliente desejar realizar uma nova locação, esta será indicada no mapa que irá lhe mostrar a vizinhança do novo filme em relação a seus principais vizinhos, que caracterizam os filmes locados mais relacionados ao qual ele deseja locar.

1.4 Organização do Documento

1.4.1 Capítulo 2: Revisão Bibliográfica

Serão apresentados os principais conceitos em que se baseia o modelo proposto por este trabalho. O capítulo inicia com uma revisão sobre sistemas de recomendação, abor-

dando as principais técnicas de recomendação, suas vantagens e desvantagens e modelos propostos para solução de problemas na recomendação. A segunda parte do capítulo apresenta um tipo especial de redes neurais artificiais: as redes auto-organizáveis ou redes SOM (*Self-organizing map*). Será abordado o modelo de redes SOM introduzido por Kohonen.

1.4.2 Capítulo 3: Modelo Proposto

Descreve a proposta central do modelo, além do algoritmo que foi desenvolvido. Neste capítulo os conceitos teóricos apresentados no capítulo 2 são utilizados de forma clara, organizada e objetiva, visando o atingimento das metas e objetivos do trabalho.

1.4.3 Capítulo 4: Configurações dos Experimentos e Análise dos Resultados

Neste capítulo serão detalhados os experimentos realizados com o modelo e os resultados obtidos para comprovar a funcionalidade do trabalho.

1.4.4 Capítulo 5: Conclusão e Trabalhos Futuros

Considerações finais, dificuldades enfrentadas durante o desenvolvimento e propostas de continuidade do trabalho.

2 *Fundamentação Teórica*

Neste capítulo estão apresentados os principais conceitos que ajudarão na compreensão do documento. Primeiramente serão abordadas as características de sistemas de recomendação estabelecidos na literatura. Também serão abordadas as principais técnicas de recomendação, apontando suas vantagens, principais problemas e soluções propostas por diversos autores a fim de solucionar essas falhas. Posteriormente serão apresentados os mapas auto-organizáveis, em especial os mapas de Kohonen, definindo sua arquitetura e o algoritmo de aprendizado.

2.1 Sistemas de Recomendação

O uso de sistemas informatizados, principalmente da internet, resulta em um grande volume de informação sendo criada e transmitida no mundo todo (3). Um estudo realizado em 2003 por pesquisadores da Universidade da Califórnia (4) estimou que cinco exabytes de informação foram criados no ano de 2002, onde a maior parte é armazenada em dispositivos magnéticos, em especial discos rígidos.

O comércio, principalmente a indústria do entretenimento, tem se beneficiado com a evolução da tecnologia. As lojas não precisam mais ter espaço físico, é possível realizar vendas e fazer negociações com segurança através da internet. A *Itunes Store* possui mais de 10 milhões de músicas disponíveis para venda através de *download*. As grandes lojas de departamentos possuem lojas virtuais para comércio eletrônico, algumas delas nem possuem lojas físicas, apenas os sítios na internet onde podem vender seus produtos. Casos de sucesso no Brasil são as Americanas.com (www.americanas.com.br) e o Submarino (www.submarino.com.br). Sem a necessidade de ter espaço físico ou prateleiras, os itens que podem ser colocados à venda são de um número superior se comparado a uma loja convencional (1). Diante de tantas possibilidades, como buscar a melhor informação? Qual produto vale a pena adquirir? Qual filme ou música escolher? Frequentemente as pessoas procuram opiniões de terceiros, como amigos e familiares que já tiveram uma ex-

periência com determinado produto ou serviço (5). Podem, ainda, procurar por resenhas em jornais e revistas, ou pedir a opinião do dono de uma livraria ou vídeo locadora. Porém, nenhum deles provavelmente vai fazer recomendações de acordo com as preferências pessoais do interessado.

Dessa falta de atenção aos reais desejos do cliente, surgiu então, a necessidade da criação de mecanismos que tenham a capacidade de filtrar ou recuperar rapidamente informação. Com o objetivo de facilitar a busca por informação foram criados mecanismos que pudessem indexar documentos na internet e, rapidamente, recuperá-los, trazendo ao usuário aquilo que ele precisa. Tais mecanismos são as ferramentas de busca tais como o *google* (www.google.com), que seleciona documentos na internet a partir de critérios que o usuário expressa através de palavras-chave. Filtrar toda a informação recebida por um usuário, raramente é uma tarefa simples e eficiente. Um dos primeiros sistemas de filtragem criados foi o *Tapestry* (6). Este sistema filtrava documentos enviados para a caixa de *emails* de um usuário. O *Tapestry* analisa, não somente o conteúdo dos textos, mas também o interesse que outros usuários tinham por esses documentos. Os idealizadores desse produto cunharam o termo "filtragem colaborativa", propondo um sistema onde a filtragem de documentos seria realizada com auxílio de grupos de pessoas com o mesmo interesse. Atualmente, os sítios de comércio eletrônico disponibilizam para seus clientes ferramentas computacionais com o objetivo de auxiliá-los no momento da compra. Essas ferramentas caracterizam os sistemas de recomendação. Tais sistemas consistem em sugerir ao usuário produtos que sejam de seu interesse ajudando-o no processo de compra, alguns sistemas ainda fornecem opiniões de outros clientes sobre aqueles produtos. Um dos maiores exemplos de sítios de comércio eletrônico que fazem bom uso de sistemas de recomendação é a *Amazon.com* (www.amazon.com).

Auxiliar o cliente mostrando produtos relacionados às suas preferências é uma forma de personalização. A personalização é uma característica do *marketing* direto. Diferente do *marketing* de massa, cujo objetivo é alcançar o maior número de pessoas através dos diversos tipos de mídia, o *marketing* direto tem seu foco no cliente individualmente. Personalizar resulta na fidelização do cliente, que é um grande diferencial entre empresas concorrentes (2), além disso, conquistar um novo cliente custa de cinco a dez vezes mais do que manter um antigo (7). Um mini-curso de sistemas de recomendação ministrado no V Encontro Nacional de Inteligência Artificial, no ano de 2005, por Eliseo Reategui e Sílvia Cazella (8) cita algumas estratégias utilizadas pelos sítios de comércio eletrônico para recomendação de produtos:

- Listas de recomendação: A loja mantém listas de produtos, como itens mais vendidos, itens que têm a melhor avaliação entre os clientes ou lista de presentes, entre outros.
- Avaliação de usuários: Consiste em se obter notas do produto por clientes que já o adquiriram, além dessa avaliação usuários podem deixar comentários sobre determinado produto (Figura 1).
- Suas recomendações: O sítio oferece alguns produtos baseado em interesses do cliente. Então pode-se ter dois tipos de recomendação: implícita, onde o sítio oferece produtos de acordo com o histórico de compras do cliente, ou explícita, onde o usuário determina quais são suas preferências (Figura 2).
- "Clientes que adquiriram um produto X, também compraram Y": O sistema de recomendação cria associações entre produtos avaliados pelo usuário para oferecer produtos relacionados ao que o cliente está adquirindo no momento (Figura 3).
- Associação por conteúdo: Este tipo de recomendação é feita baseado no conteúdo de determinado item. Por exemplo: os livros: Redes de Computadores e Programação em Java são freqüentemente vendidos em conjunto.

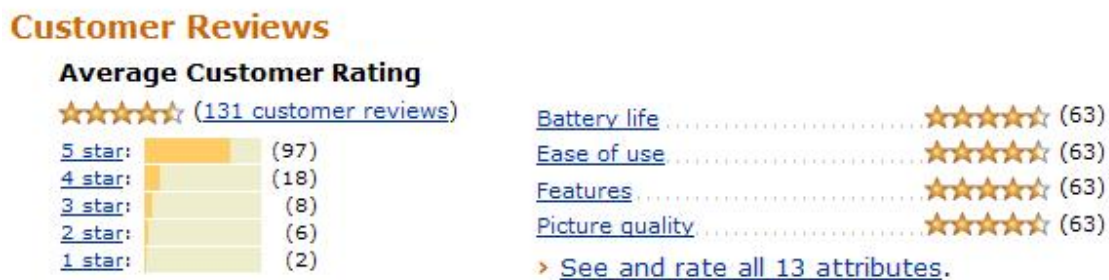


Figura 1: Avaliações de usuários no sítio da Amazon.com para determinado produto.

2.1.1 Técnicas de recomendação

Diversas técnicas que pudessem identificar padrões de comportamento e filtragem de informação foram definidas na literatura com o objetivo de obter recomendações e personalização para o usuário. As três técnicas mais utilizadas em sistemas de recomendação são a filtragem baseada em conteúdo, a filtragem colaborativa e a filtragem híbrida, que procura conciliar as vantagens de duas ou mais técnicas atacando seus principais problemas.



Figura 2: Recomendações da Amazon.com de acordo com o histórico do cliente.



Figura 3: Associação de produtos por clientes na Amazon.com

2.1.1.1 Filtragem baseada em conteúdo

A filtragem baseada em conteúdo tem suas raízes no processo chamado de recuperação de informação, onde o usuário apresenta ao sistema um formulário e recebe, como resultado, documentos associados a esses critérios (9). O principal objetivo da recuperação de informação é encontrar documentos que correspondam a determinado critério de busca (10). Em um sistema de recuperação de informação, o usuário fornece ao sistema palavras-chave que representam seus interesses ou necessidades atuais na procura por informação. O sistema então, realiza uma busca por essas palavras em documentos armazenados numa base e retorna os documentos mais relevantes para os critérios informados.

A filtragem de informação inspeciona todo novo item que é recebido selecionando aqueles relacionados a seus interesses. Assim, a filtragem de informação, mantém o perfil do usuário atualizado e demonstra interesses de longo prazo, pois continua realizando a seleção de documentos de acordo com as preferências identificadas no perfil. Por outro lado, a recuperação de informação, que seleciona documentos de acordo com um critério de pesquisa, representa interesses de curto prazo (11), ou seja, interesses momentâneos traduzidos pelos critérios de busca do usuário. Os atuais sistemas de recomendação têm sua origem nos sistemas de filtragem de informação.

Em 1982, Peter Denning (12) já apontava para o problema do volume de informação produzida e transmitida através de sistemas computacionais e a facilidade de compartilhamento de informações pela internet, sendo necessária uma atenção maior para o processo de controlar e filtrar toda essa informação. Filtragem de informação e filtragem baseada em conteúdo são termos semelhantes e ambos possuem o mesmo objetivo: filtrar itens através da análise do seu conteúdo (10).

Na filtragem baseada em conteúdo, as recomendações são feitas apenas baseadas em um perfil do usuário previamente construído. Esse perfil é individual e determinado a partir da análise do conteúdo de itens que o usuário qualificou ou mostrou algum interesse no passado (9). Quando o usuário de um sítio de comércio eletrônico, por exemplo, entra na página da loja e revela suas necessidades através de palavras-chave na ferramenta de busca, ele está realizando uma recuperação de informação, pois o sistema apresenta produtos que satisfaçam simplesmente os critérios apresentados no momento. Quando este mesmo sítio armazena o perfil do cliente e apresenta produtos semelhantes aos que este usuário mostrou interesse no passado, caracteriza uma filtragem de informação. Uma das técnicas mais populares para representação dos itens em sistemas de filtragem baseada em conteúdo é a TF-IDF (Term-frequency Inverse-Document-Frequency). Esta técnica realiza comparação e cálculo de similaridade a partir da frequência de ocorrência de palavras-chave nos textos (2). Para criação do perfil do usuário, normalmente são utilizadas técnicas de computação inteligente, que podem extrair informações do comportamento de determinado usuário, por exemplo, algoritmos de classificação podem identificar e fazer a divisão entre itens que o usuário gosta e itens que ele não gosta (10). O *feedback* é muito importante na fase de aprendizado, podendo ser explícito, quando o usuário qualifica o item que lhe foi oferecido, ou implícito, quando o usuário demonstra interesse no item finalizando com a compra (3). Exemplos de sistemas que utilizam a filtragem baseada em conteúdo são o NewsWeeder (13) e o InfoFinder (14).

Dentre as vantagens da utilização de filtragem baseada em conteúdo, podem-se citar: (i) não é necessário que um novo item tenha algum tipo de qualificação, bastando apenas que este seja semelhante ao perfil do usuário e (ii) por avaliar a semelhança entre todos os itens, a filtragem baseada em conteúdo não se restringe a itens já avaliados por outros usuários. Os principais problemas apresentados pela filtragem baseada em conteúdo são: (i) as representações de conteúdo de um item não conseguem capturar alguns aspectos, como qualidade de um texto, no caso sistemas de recomendação de documentos, ou informações de multimídia, no caso de sistemas que recomendam páginas da internet (9); (ii) o segundo ponto negativo da filtragem baseada em conteúdo é a superespecialização do

sistema. A diversidade de itens recomendados fica restrita à similaridade de itens avaliados pelo usuário, sem conseguir recomendar itens não relacionados que poderiam ser de interesse do cliente.

2.1.1.2 Filtragem colaborativa

A filtragem colaborativa é a técnica de recomendação mais comum e a mais largamente utilizada (15). A abordagem da filtragem colaborativa consiste em recomendar itens baseados na similaridade entre usuários, diferentemente da filtragem baseada em conteúdo, que realiza similaridade entre itens. Basicamente, é formado um conjunto de usuários denominados "vizinhos mais próximos", que possuem itens classificados de forma semelhante ao usuário alvo. Os itens que serão recomendados são itens que foram bem qualificados por esse grupo de usuários, partindo do pressuposto que esse mesmo grupo possui preferências em comum com o usuário alvo. A priori, neste tipo de técnica nenhuma informação sobre os itens em si é conhecida, as recomendações são baseadas na semelhança entre os usuários (9). Segundo descreve Burke (15): "O perfil de um usuário num sistema colaborativo consiste em um vetor de itens e suas qualificações, que é aperfeiçoado na medida em que o usuário interage com o sistema ao longo do tempo."

Um dos primeiros sistemas a utilizar a filtragem colaborativa foi o *Tapestry* (6), mencionado anteriormente neste trabalho no início da seção 2.1. A idéia do *Tapestry* é fazer com que as pessoas colaborem entre si para realizar filtrações descrevendo suas percepções referentes a documentos que receberam. Essas percepções são armazenadas em comentários ou anotações que podem ser consultadas por qualquer outro usuário. Uma forma de utilização desse sistema seria um usuário buscar documentos que outro determinado usuário aprovou. Os principais representantes de sistemas que utilizam filtragem colaborativa, além do *Tapestry*, são o *GroupLens* (16), *Ringo* (17) e *Bellcore* (18).

A filtragem colaborativa também possui vantagens e problemas identificados por diversos especialistas na literatura. Algumas vantagens deste tipo de técnica são:

- Como as recomendações são baseadas em avaliações de outros usuários é possível tratar diferentes tipos de conteúdo, não somente documentos textuais.
- A qualidade das recomendações é superior se comparado a sistemas que utilizam filtragem baseada em conteúdo. A justificativa está no fato de sistemas automatizados não serem eficientes em determinar a qualidade de um produto analisando apenas seu conteúdo. A qualidade de um item está implícita na avaliação de outros

usuários, sendo possível produzir recomendações de melhor qualidade.

Algumas das desvantagens da filtragem colaborativa são:

- Se um novo item é introduzido no sistema, não será recomendado até que um dos usuários o avalie. O que traz outra consequência: se o número de usuários do sistema é reduzido em relação à quantidade de itens, ocorre um problema de cobertura, ou seja, a coleção de itens que podem ser recomendados será resumida.
- Outro problema em sistemas utilizando filtragem colaborativa ocorre quando há um usuário cujas preferências diferem dos demais usuários, conhecidos como "ovelhas-negras". Neste caso, não há vizinhos semelhantes ao usuário ativo, ocasionando recomendações que não correspondem às preferências desse usuário.

2.1.1.3 Filtragem híbrida

A abordagem da filtragem híbrida consiste em associar duas ou mais técnicas de recomendação com o intuito de unir suas vantagens e, simultaneamente, tratar os principais problemas inerentes a cada uma delas. Observando um resumo das vantagens e desvantagens das filtragens colaborativa e baseada em conteúdo, apresentado na Tabela 1, pode-se perceber que essas duas abordagens são complementares. Fato que motivou a criação de sistemas híbridos (9).

Tabela 1: Vantagens e desvantagens das filtragens baseada em conteúdo e colaborativa

Filtragem colaborativa		Filtragem baseada em conteúdo	
Vantagens	Desvantagens	Vantagens	Desvantagens
É possível tratar diferentes tipos de conteúdo	Problema da primeira avaliação	Não é necessário que um novo item tenha algum tipo de qualificação	Trata somente conteúdo textual
A qualidade de um item está implícita na avaliação	Usuário "ovelha-negra"	Não se restringe a itens já avaliados por outros usuários	Superespecialização do sistema

Burke (15) realizou um estudo sobre sistemas de recomendação híbridos. Neste artigo ele aponta sete combinações possíveis entre técnicas de recomendação:

1. Ponderado: A avaliação de um item é a combinação do resultado de todas as técnicas presentes no sistema. O sistema *P-Tango* (19) é um exemplo de uso desta técnica.

A proposta do *P-Tango* é combinar as recomendações da filtragem baseada em conteúdo e da filtragem colaborativa através de médias ponderadas. Os pesos são ajustados de acordo com o peso de cada recomendação, isto é, ao passo que o número de usuários e avaliações de um determinado item cresce, a filtragem colaborativa tem seu peso aumentado;

2. Alternado: O sistema se utiliza de um critério para alternar entre as técnicas de recomendação. Por exemplo, se uma filtragem baseada em conteúdo não retorna resultados satisfatórios, então a recomendação é feita com a filtragem colaborativa;
3. Misto: Neste tipo de combinação, o sistema apresenta recomendações de diferentes técnicas numa mesma listagem. O sistema PTV (20) utiliza a filtragem baseada em conteúdo baseado nas descrições de programas de TV e a filtragem colaborativa baseado na preferência de outros usuários, o resultado das recomendações é uma combinação dos resultados das duas filtragens;
4. Combinação de características: Trata informações de um tipo de recomendação como simples características que são utilizadas por outra técnica para retornar recomendações mais significativas. (21) apresenta um sistema de recomendação de filmes onde as informações da filtragem colaborativa (denominadas, neste artigo, características colaborativas) são influenciadas pelo conteúdo agregado aos filmes. Desta forma, é possível agrupar usuários semelhantes e que gostem do mesmo gênero de filmes;
5. Cascata: Nesta combinação, uma técnica, primeiramente, produz suas recomendações e uma segunda técnica é usada para refinar os resultados da anterior;
6. Acréscimo de característica: Primeiramente uma técnica é aplicada para gerar classificações para um item, então estas informações são incorporadas na técnica seguinte;
7. *Meta-level*: Esta abordagem utiliza o modelo criado por uma técnica como entrada para outra técnica. No acréscimo de característica, um modelo gera características para serem usadas como entrada para outro modelo, na abordagem *meta-level* o modelo completo é utilizado como entrada para outro modelo.

2.1.2 Confiança em sistemas de recomendação

O objetivo central dos sistemas de recomendação é levar o cliente de uma dada loja a fazer melhores compras, tornando o mesmo mais satisfeito com o serviço. Para que

o processo de recomendação seja realmente eficaz é importante convencer o usuário de que as recomendações são relevantes e que estão relacionadas à satisfação do cliente (7). Prover argumentos e explicações sobre os dados que geraram a recomendação agrega maior confiança ao sistema. A transparência em sistemas de recomendação melhora o entendimento pelo cliente e este se mostra mais propenso a aceitar as recomendações (10).

2.2 Mapas auto-organizáveis

O principal objetivo das redes neurais artificiais (RNAs) é aprender com o ambiente em que estão inseridas para adquirirem a capacidade de adaptação às mudanças nesse ambiente. O tipo de aprendizado das redes neurais pode ser dividido em duas classes: aprendizado supervisionado e aprendizado não-supervisionado. O método mais comumente utilizado é o aprendizado supervisionado, onde, além da entrada, as saídas desejadas para a rede são fornecidas por um supervisor externo (22). A tarefa então é fazer o ajuste de parâmetros de forma que a rede possa representar novas entradas corretamente. O método de aprendizado supervisionado se limita, porém, a aplicações onde tanto os dados de entrada como as respostas desejadas são conhecidos. Algumas aplicações necessitam que o treinamento seja realizado sem um supervisor para subsidiar o treinamento da rede, todo o conhecimento deve ser adquirido somente através das entradas. Diversos algoritmos foram desenvolvidos para treinar redes sem supervisão, este trabalho irá tratar exclusivamente de um deles: os modelos *self-organizing* (ou auto-organizáveis).

2.2.1 Visão Geral

Os mapas *self-organizing*, SOM (*Self-organizing maps*) foram propostas por Teuvo Kohonen (23). A estrutura básica dessas redes é formada por uma camada de entrada, constituída de padrões fonte, e uma camada de saída (geralmente uma grade 2D). Não há camadas escondidas ("hidden-layer") no modelo de mapas auto-organizáveis como em outros tipos de redes neurais. A inspiração para o desenvolvimento das redes SOM está no mapa topologicamente organizado do córtex cerebral de animais mais desenvolvidos. O córtex cerebral humano é dividido de acordo com funções específicas, como fala e visão (Figura 4). Desta forma, neurônios que lidam com o mesmo tipo de estímulo estão localizados próximos uns dos outros. O modelo de Kohonen consiste em um mapa, onde os nodos ou neurônios artificiais que estão topologicamente próximos respondem a estímulos

semelhantes, ou seja, padrões de entrada que possuem similaridade entre si.



Figura 4: Córtex cerebral humano

Na arquitetura de uma rede SOM, os nodos são dispostos em uma grade ou reticulado, geralmente bidimensional ou unidimensional, com raras exceções, há redes tridimensionais ou n-dimensionais. No modelo bidimensional, os neurônios estão organizados em linhas e colunas, como mostra a Figura 5. Cada nodo possui um conjunto de pesos que representam as sinapses do neurônio biológico, esses pesos são ajustados de maneira que o nodo represente um dado padrão de entrada. Os nodos de uma rede SOM funcionam como um extrator de características, quanto mais o vetor de pesos de um neurônio for semelhante a um padrão de entrada, maior será sua saída e mais representativo este nodo será para a entrada (22).

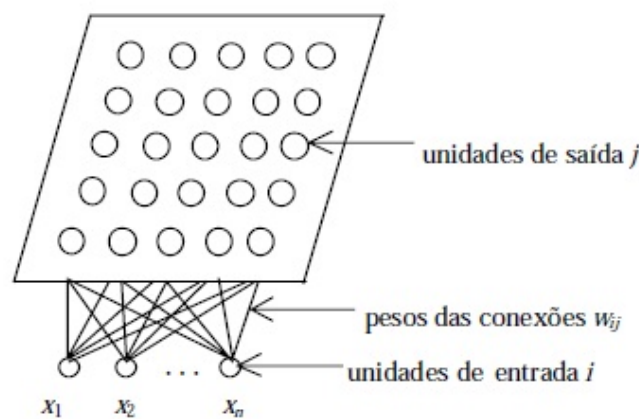


Figura 5: Rede SOM bidimensional

As redes SOM utilizam um processo de aprendizado competitivo, no qual os neurônios da camada de saída competem entre si para representar um dado padrão de entrada, assim, apenas um neurônio de saída ou neurônio por grupo estará ativo a qualquer instante

de tempo. O neurônio que se sobressai entre os outros para representar a entrada é chamado de vencedor e a competição é chamada de *winner-takes-all*, o vencedor leva tudo. Para implementar esta competição são normalmente utilizadas conexões laterais inibitórias entre os neurônios de saída. O modelo para esse tipo de conexão também provém das células do córtex cerebral, onde a ordenação topológica dos neurônios se dá graças ao *feedback* lateral entre as células. Em RNAs este *feedback* é modelado por uma função chamada chapéu mexicano. Segundo esta função, as interações laterais entre os neurônios podem ser divididas em três regiões distintas, como mostrado na Figura 6: (1) área excitatória, vizinhos que estão mais próximos ao neurônio atual; (2) área inibitória, vizinhos que estão fora da área anterior, mas incluídos numa segunda área; e (3) área levemente excitatória, que rodeia a área inibitória, esta terceira área geralmente é ignorada.

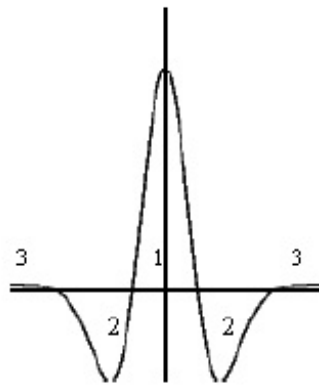


Figura 6: Função chapéu mexicano

Para simular o efeito da função chapéu mexicano, a rede SOM utiliza o conceito de vizinhança topológica dos neurônios vencedores. Quando um neurônio vence a competição e é o escolhido para representar o padrão de entrada, ele tem seus pesos ajustados de forma a se aproximar mais da entrada, com o conceito de vizinhos topológicos, além do neurônio vencedor ter seus pesos ajustados, os neurônios localizados na vizinhança também têm seus pesos ajustados.

2.2.2 Treinamento

Como afirmado anteriormente, o treinamento de redes SOM é competitivo e não-supervisionado. Primeiramente os pesos dos neurônios do mapa são inicializados com valores aleatórios, que serão ajustados ao longo do algoritmo de aprendizado, de forma que se aproximem dos padrões de entrada. Em seguida é apresentado um padrão p

à rede, neste momento a rede define o neurônio que melhor representa esta entrada, caracterizando o neurônio vencedor. Para a escolha do neurônio vencedor é definida uma função de ativação que é baseada na distância entre o peso do neurônio e o vetor de entrada. A função de ativação mais conveniente para a rede SOM é baseada na distância euclidiana (24), apresentada na equação 2.1:

$$y_j = \sum_{i=1}^n \|x_i - w_{ji}\| \quad (2.1)$$

onde y_j representa a saída do neurônio j , x é o vetor de entrada e w_{ji} é o peso do neurônio j associado ao elemento de entrada x_i .

O neurônio que possui a menor distância é escolhido como o vencedor e irá representar o padrão de entrada. Após essa escolha dá-se início ao processo de atualização dos pesos. Nesta fase o neurônio vencedor e os vizinhos definidos pelo raio ou área de vizinhança atualizam seus pesos. A fim de implementar a interação lateral, é definida uma região de vizinhança N_c , tendo como centro o neurônio c , estabelecido como vencedor pela função de ativação. Todos os neurônios internos a essa vizinhança terão os pesos atualizados, enquanto neurônios fora do limite serão deixados intactos. Inicialmente, o valor do raio ou tamanho de N_c deve ser alto, abrangendo, aproximadamente, quase todo o mapa e depois devendo diminuir monotonicamente no tempo (24). Tal valor pode, ao final do processo, abranger apenas o neurônio central ($N_c = \{c\}$), como se pode observar na Figura 7. A equação 2.2 mostra como são atualizados os pesos do neurônio vencedor e dos neurônios vizinhos.

$$w_{ji}(t+1) = \begin{cases} w_{ji}(t) + \alpha(t)(x_i(t) - w_{ji}(t)), & \text{se } j \in N_c(t) \\ w_{ji}(t), & \text{se } j \notin N_c(t) \end{cases} \quad (2.2)$$

onde $\alpha(t)$ é o valor da taxa de aprendizado $0 < \alpha(t) < 1$, que decresce no tempo.

Como alternativa pode ser introduzida uma função de vizinhança do neurônio vencedor, definido pela seguinte equação, com r_c e r_j como as coordenadas dos neurônios c (neurônio central ou vencedor) e j , respectivamente:

$$h_{ci}(t) = h_0 \exp(-\|r_i - r_c\|^2 / \sigma^2) \quad (2.3)$$

onde $h_0 = h_0(t)$ e $\sigma = \sigma(t)$ são funções que devem decrescer no tempo. O parâmetro $\sigma(t)$ corresponde ao raio de $N_c(t)$. O ajuste de pesos passa então a ser calculado desta forma:

$$w_{ji}(t+1) = w_{ji}(t) + h_{ci}(x_i(t) - w_{ji}(t)) \quad (2.4)$$

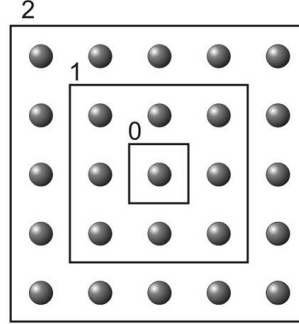


Figura 7: Exemplo de vizinhança, onde o instante 2 é menor que o instante 1, que por sua vez é menor que o instante 0

Segundo estudos e experiências na escolha dos parâmetros, Kohonen (24) recomenda que o valor inicial de $\alpha(t)$ (taxa de aprendizado da rede) deve estar próximo de 1 e decair monotonicamente durante os primeiros 1000 ciclos da fase de aprendizado, porém mantendo o valor acima de 0,1. A regra para o decréscimo de $\alpha(t)$ pode ser uma função linear, exponencial ou inversamente proporcional a t , por exemplo: $\alpha(t) = 0.9(1 - t/1000)$. É durante esta fase inicial do treinamento que ocorre a fase de ordenação da rede. Nas fases seguintes ocorre o ajuste fino da rede, chamado de fase de convergência. O número de ciclos da fase de aprendizado deve ser razoavelmente grande. Uma regra empírica é que este número deva ser 500 vezes maior que o número de neurônios na rede. O tamanho da vizinhança de um neurônio não pode ser muito pequeno inicialmente, pois o mapa não teria uma boa ordenação global. A princípio o raio ou tamanho inicial da vizinhança pode ser maior que a metade do tamanho do mapa. A Figura 8 apresenta um resumo da execução do algoritmo de aprendizado das redes SOM e a Figura 9 apresenta o fluxograma do funcionamento deste processo.

2.2.3 Aplicações

As redes SOM foram inicialmente idealizadas para identificar relações não-lineares entre dados apresentados em mais de duas dimensões. Com a popularidade do algoritmo, diversas aplicações foram desenvolvidas e problemas solucionados com o uso dos mapas auto-organizáveis. Análise de dados exploratórios, reconhecimento de padrões e reconhecimento e análise de fala estão entre as áreas de aplicação de redes SOM. Um exemplo de aplicação desenvolvida pelo próprio Kohonen na área de reconhecimento de fala é o

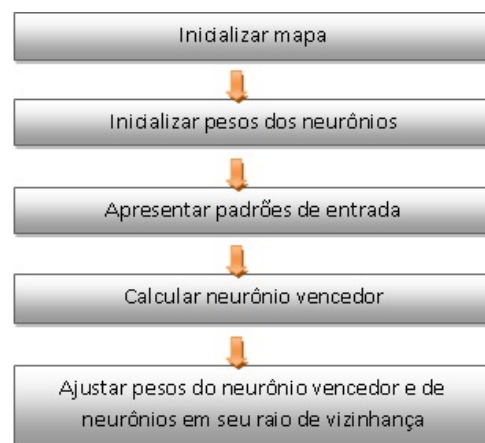


Figura 8: Resumo do algoritmo de aprendizado SOM

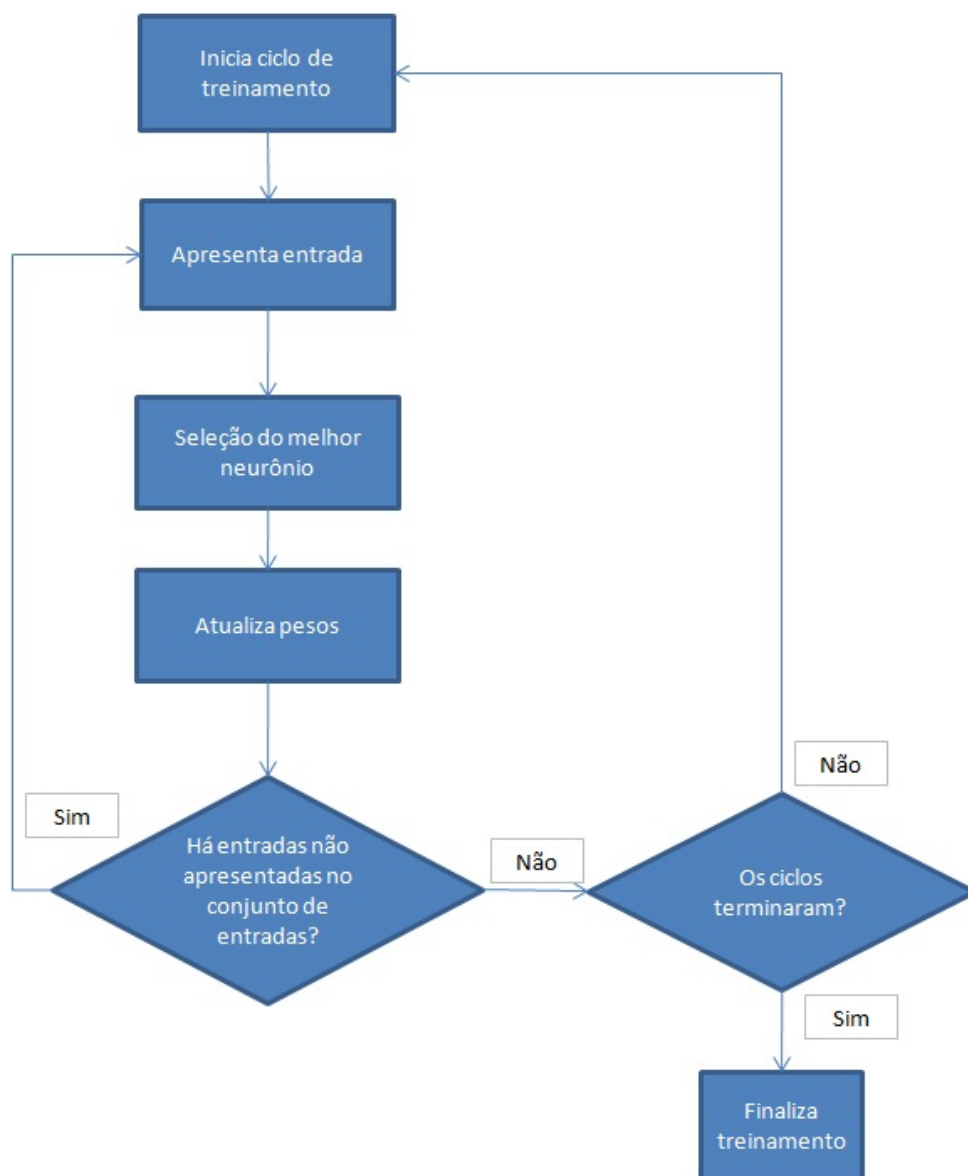


Figura 9: Fluxograma do algoritmo de aprendizado das redes SOM

neural phonetic typewriter (25). Outra aplicação que merece destaque é o WEBSOM (26), no qual Kohonen também participou do desenvolvimento. WEBSOM é um método para organização de grandes coleções de documentos de texto. Baseado em redes SOM, o WEBSOM organiza documentos num reticulado bidimensional de forma que documentos relacionados apareçam próximos.

2.2.4 Trabalhos relacionados

Existem algumas aplicações na literatura de sistemas de recomendação que utilizam redes SOM. Esta seção se dedica ao comentário de duas das mais importantes.

Graef e Schafer (27) descrevem duas abordagens baseadas em filtragem colaborativa usando redes SOM e ART2. Eles investigaram o desempenho dos dois modelos avaliando tempo de resposta, qualidade das predições e adaptatividade. A abordagem utilizando redes SOM consiste em agrupar usuários no mapa baseado na semelhança de suas avaliações sobre filmes.

Um sistema de recomendação híbrido combinando filtragem colaborativa com SOM é apresentado por Lee (28). Essa abordagem utiliza redes SOM para criar agrupamentos de usuários de acordo com informações demográficas, como sexo e idade e preferências por diversos gêneros de filmes. A idéia é aplicar filtragem colaborativa dentro de agrupamento ao qual o usuário pertence.

3 *Modelo proposto*

Este trabalho visa apresentar um modelo para recomendação de filmes para clientes no ambiente de vídeo locadoras. O modelo proposto utiliza as redes SOM como descritas por Kohonen (23) para aprender o perfil do usuário e auxiliá-lo na escolha do filme.

Após a descrição dos conceitos e algoritmos apresentados no capítulo 2, será apresentada neste capítulo a contribuição do modelo proposto.

3.1 Visão Geral

O modelo para recomendação desenvolvido leva em consideração a limitação de vídeo locadoras no que tange aos conceitos gerais de sistemas de recomendação comuns. As informações que estão disponíveis num ambiente de locadora de filmes consideram somente o cliente como indivíduo e não um grupo de clientes que podem contribuir juntamente para gerar recomendações. Portanto não há o conceito de avaliação de um item pelo usuário. Nos sistemas de locadoras comuns não existe um mecanismo onde o cliente possa dar sua nota para um filme de forma que outros clientes possam acompanhar essas avaliações e ter um parâmetro para a escolha de determinado título. Existe apenas a opinião "de boca" de clientes que queiram expressar ou quando o funcionário da loja educadamente questiona sobre a satisfação do cliente com relação a determinado filme. Nenhuma informação é armazenada num banco de dados, nem é realizado um levantamento das opiniões de diversos clientes, com o propósito de identificar os melhores filmes.

A proposta deste trabalho é desenvolver um sistema que, baseado no histórico de locações de um cliente, ele possa auxiliar a condução do processo de uma escolha satisfatória no momento de locar um novo título. O sistema se baseia somente no conteúdo de informação acerca dos filmes que já foram locados por um determinado cliente, informações essas que podem ser facilmente obtidas no banco de dados da locadora. Como mencionado anteriormente, não existem avaliações de clientes sobre os filmes nem parâmetros

que determinem as preferências desses clientes. Portanto, o modelo proposto não ambiciona gerar uma lista de recomendações diretas com títulos para o cliente, apenas auxiliá-lo de forma a realizar uma escolha consciente baseado nas informações contidas nos filmes. O modelo utiliza redes auto-organizáveis, pois este algoritmo desempenha bem a função de agrupamento e vizinhança, para que o cliente tenha uma percepção melhor da relação entre os filmes que já assistiu com os que deseja locar atualmente. O objetivo específico do trabalho é realizar uma prova de conceito para o uso de redes SOM na recomendação de filmes a ser implementado em vídeo locadoras como forma de personalização da loja para o cliente, tornando mais prazeroso e simplificado o ato da escolha do filme a ser locado. A personalização, como comentado anteriormente, torna o cliente mais satisfeito e mais assíduo.

O modelo proposto consiste em uma rede SOM que, quando treinada, é composta por filmes contidos no histórico de um cliente (topologicamente organizados). Propomos então, um mapa SOM para cada cliente; e não um mapa só para toda a locadora. Essa decisão de projeto respeita a personalização, objetivo principal deste trabalho. A rede irá distribuir os títulos no mapa bidimensional, agrupando-os de acordo com semelhanças entre informações fornecidas sobre os filmes. Nesta fase ocorre o aprendizado do comportamento do cliente com relação às escolhas de filmes.

Após o treinamento da rede, os filmes locados por um cliente estarão distribuídos no mapa. O cliente deve, então selecionar um filme do acervo da locadora e apresentar ao seu mapa individual. A rede irá calcular a posição deste novo filme e irá mostrar ao usuário três filmes que estejam próximos ao primeiro, determinando que há semelhanças entre esses quatro títulos. Esta é uma forma de auxiliar o cliente, pois ele pode avaliar se irá gostar ou não do filme, partindo da satisfação que teve ao assistir os outros três. A Figura 10 mostra o fluxo básico da execução do sistema.

A ferramenta foi modelada de forma a se integrar ao sistema de informação da vídeo locadora com o objetivo de ser um diferencial da locadora para corresponder à demanda dos clientes na busca de informações que os auxiliem no processo de escolha e locação de novos filmes. O projeto proposto, então, não pretende realizar o gerenciamento do estabelecimento como um todo.

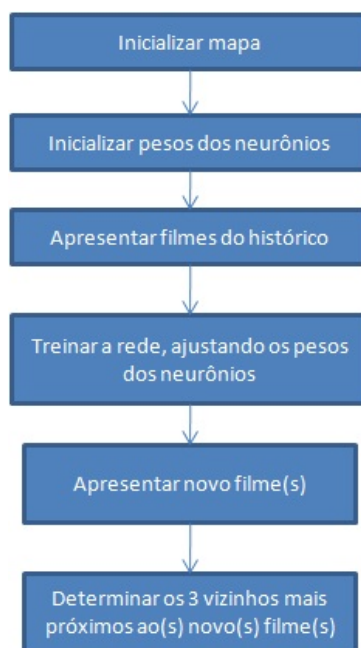


Figura 10: Fluxo de execução do modelo proposto

3.2 Implementação do modelo

A implementação da rede SOM utilizada no modelo deste trabalho foi feita baseada no algoritmo original proposto por Kohonen, como descrito no Capítulo 2. A linguagem utilizada para desenvolvimento foi C# com o Microsoft® Visual Studio® 2008 como IDE (*Integrated Development Environment* ou Ambiente Integrado de Desenvolvimento) de programação. A escolha desta tecnologia se justifica por ser uma linguagem de fácil implementação e ser mais largamente utilizada pelo autor deste trabalho. A Figura 11 mostra os diagramas de classes do modelo implementado. MapaSOM é uma classe que comporta todos os neurônios da rede e recebe o conjunto de entradas para ser apresentado durante o treinamento. Cada neurônio possui coordenadas x e y , determinando sua posição no reticulado do mapa e o vetor de pesos que são inicializados aleatoriamente e ajustados durante a fase de aprendizado. Cada padrão de entrada tem um conjunto de características que são extraídos do banco de dados de filmes, além de um objeto neurônio que caracteriza o neurônio vencedor para representar o padrão.

A Figura 12 mostra o diagrama de sequência do programa durante a fase de treinamento. O treinamento inicia criando um mapa com um determinado número de neurônios, cujos pesos são inicializados com valores aleatórios. Após a fase de inicialização do mapa, tem início o algoritmo de aprendizado propriamente dito. Então são apresentados os padrões de entrada que representam os filmes presentes no histórico do cliente e, para cada

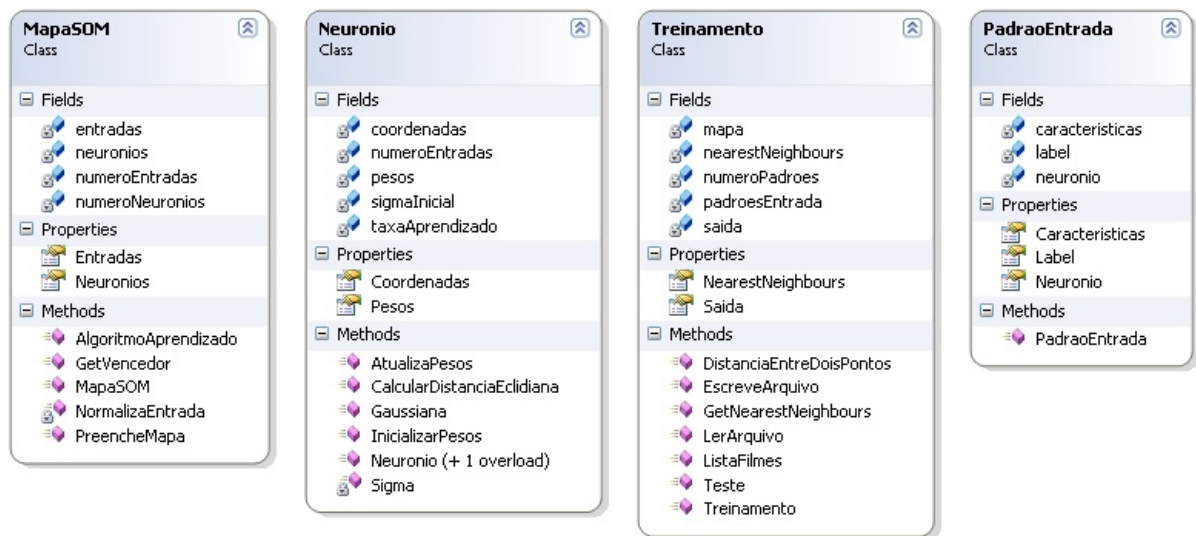


Figura 11: Diagrama de classes do modelo

um deles, é selecionado o neurônio que melhor representa uma dada entrada. Este neurônio tem seus pesos atualizados de forma a melhor refletir as características da entrada. Os neurônios contidos no raio de vizinhança do neurônio vencedor também têm seus pesos atualizados. Assim que todos os padrões de entrada, ou seja, todos os títulos do histórico, forem apresentados, é iniciado um novo ciclo de treinamento. Quando o número de ciclos atingir um determinado limite, o treinamento é encerrado. O valor do limite de ciclos segue o valor proposto por Kohonen, como mencionado na seção 2.2.2.

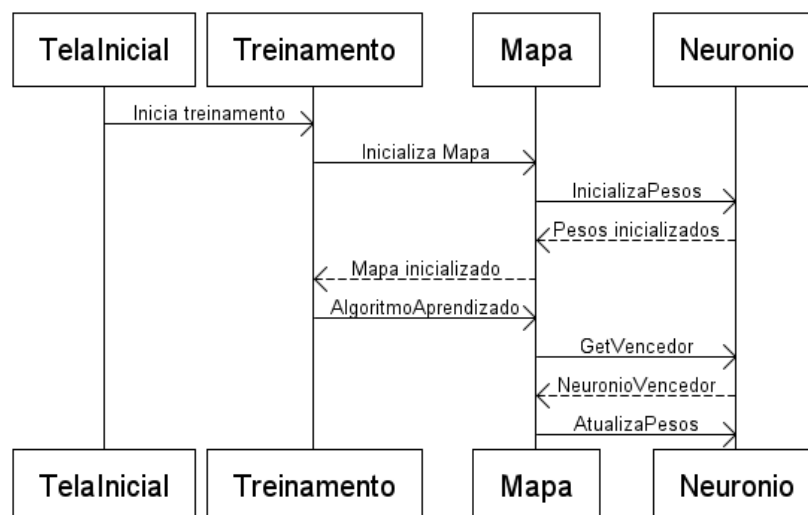


Figura 12: Diagrama de sequência da fase de treinamento

3.3 Funcionamento do modelo

O algoritmo inicia obtendo os padrões de entrada a serem utilizados na fase de treinamento, que são os filmes do histórico de locações do cliente. A ferramenta extrai os títulos dos filmes locados pelo cliente e busca no acervo da locadora as informações complementares para cada título. Essas informações serão utilizadas como vetor de entrada para cada padrão de entrada durante a fase de aprendizado da rede. Logo em seguida é determinado o tamanho do mapa, ou seja, o número de neurônios presentes na rede. O vetor de pesos de cada um desses neurônios é inicializado com valores aleatórios e é determinado o raio de vizinhança inicial. Empiricamente, um valor inicial adequado para a vizinhança é de 60% do tamanho total do mapa.

Após a fase de inicialização do mapa, segue-se a fase de aprendizado ou treinamento. A fase de treinamento inicia ao serem apresentados à rede os padrões de entrada. Cada padrão de entrada possui um vetor de características que identifica unicamente um filme. Então é calculado o neurônio que mais se aproxima de cada padrão, denominado como neurônio vencedor. O cálculo para determinar o neurônio vencedor segue a função da distância euclidiana como descrito na seção 2.2.2 pela equação 2.1. Este cálculo é realizado através da distância euclidiana entre o vetor de características de um padrão de entrada e o vetor de pesos de um neurônio. Quanto mais similar um neurônio for de uma entrada menor será a distância entre os dois, caracterizando o neurônio vencedor.

Os pesos do neurônio selecionado são atualizados, assim como os pesos de neurônios em seu raio de vizinhança. Como dito anteriormente o raio de vizinhança inicial é igual a 60% do tamanho do mapa, esse raio é decrescido monotonicamente segundo a equação 2.3 mostrada na seção 2.2.2. Esses passos são repetidos até que o número determinado de ciclos seja alcançado. No final da fase de treinamento todos os filmes presentes no histórico do cliente estão distribuídos topologicamente no mapa que representa, então, o perfil do usuário.

Com a rede treinada, inicia-se a fase de teste. Neste passo serão mostrados novos títulos de filmes à rede, representando filmes que o cliente deseja alugar. Ao ser apresentado a um novo padrão, a rede irá calcular o neurônio que melhor representa este padrão, da mesma maneira que é realizado na fase de aprendizado. A diferença aqui está em que os pesos do neurônio vencedor não serão mais ajustados, ele é apenas selecionado como representante daquela entrada no mapa topológico do perfil do cliente. O ajuste de pesos é que caracteriza o aprendizado da rede, portanto nesta fase não é realizado o cálculo de

ajuste e de vizinhança.

Após o cálculo do neurônio vencedor para o novo filme, o algoritmo irá determinar quais são os três neurônios mais próximos daquele. Para isso é calculada a distância entre dois pontos no plano bidimensional. Os três neurônios mais próximos do vencedor representam padrões de entrada que são os filmes que mais se assemelham ao filme representado pelo neurônio vencedor. O sistema então auxilia o cliente na decisão mostrando que o filme que ele deseja locar tem relação de semelhança com outros três que ele já locou no passado. Dependendo da satisfação do usuário com os três filmes que foram mostrados, ele decide se irá mesmo locar o filme escolhido ou não. Prover a informação da relação entre os filmes agrega confiança ao sistema, pois o cliente saberá a origem das recomendações apresentadas.

Para visualização do mapa, a ferramenta cria um gráfico bidimensional plotando os filmes do histórico e os novos filmes que o cliente deseja locar. As Figuras 13 e 14 mostram exemplos desse gráfico para o cliente 1 e cliente 2, respectivamente. Cada ponto do gráfico representa um neurônio contido no mapa que foi construído na fase de treinamento. Os pontos de cor laranja representam os títulos presentes no histórico do cliente, os pontos de cor vermelha são os novos filmes que o cliente deseja locar e os de cor azul são pontos que não representam nenhum filme. Navegando pelo mapa o cliente pode aproximar o *mouse* a um dos pontos e será mostrado um *tooltip* (legenda) contendo o título do filme que o ponto representa, desta forma o cliente pode procurar os filmes que estão mais próximos dos novos filmes.

Os gráficos apresentados foram criados utilizando a ferramenta *Visifire*, que cria gráficos utilizando a tecnologia WPF (Windows Presentation Foundation) presente na plataforma .NET da Microsoft®. O WPF é um subsistema gráfico do .NET Framework, ou seja é uma tecnologia que permite a criação de interfaces de apresentação para aplicativos desenvolvidos sob a plataforma .NET.

A nível de experimentos e avaliação de resultados, foram gerados arquivos de texto contendo as informações dos novos filmes e dos três filmes que mais se assemelham a eles. Estes experimentos serão detalhados no Capítulo 4.

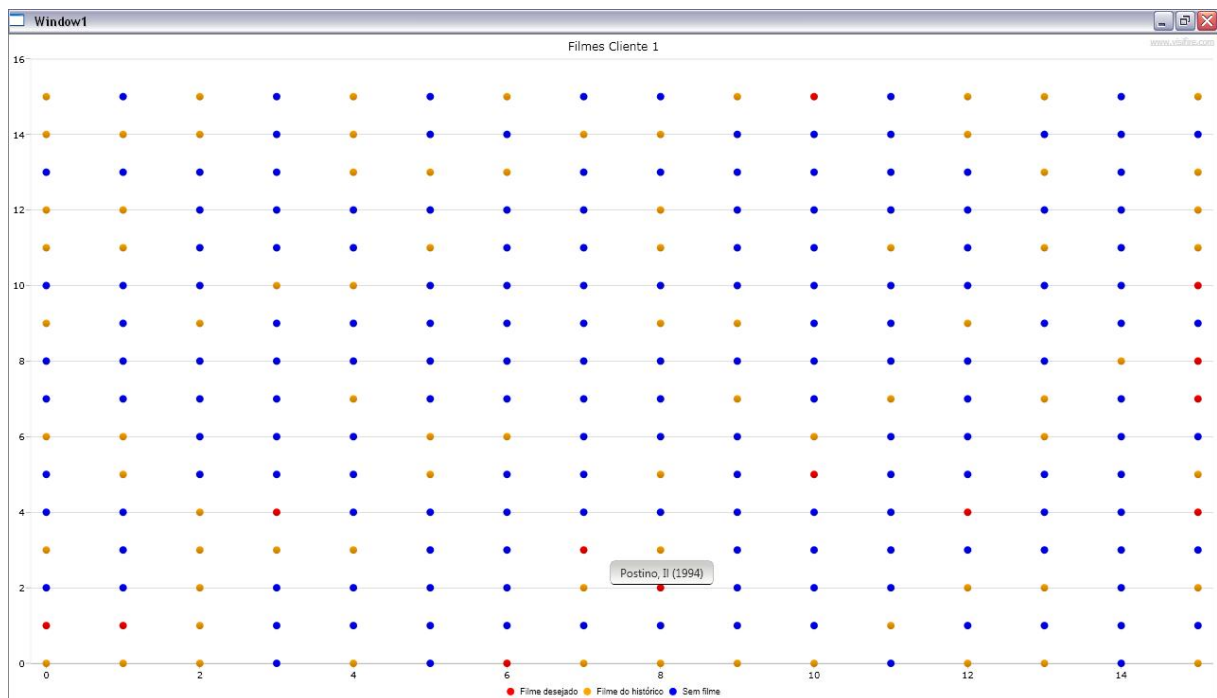


Figura 13: Mapa de filmes do cliente 1



Figura 14: Mapa de filmes do cliente 2

4 *Experimentos e Análise de Resultados*

Este capítulo tem como objetivo descrever os experimentos realizados e resultados obtidos a partir da implementação do modelo descrito no Capítulo 3.

4.1 Base de dados

Como mencionado no capítulo anterior, o objetivo deste trabalho é auxiliar clientes na escolha de filmes em uma vídeo locadora. Foi desenvolvida uma rede auto-organizável, segundo o modelo de Teuvo Kohonen (24), com o fim de aprender o comportamento do usuário e poder guiá-lo em sua escolha. Para a realização dos testes foi utilizada uma base de dados real, extraindo as características que mais se adequam ao modelo proposto.

Inicialmente estava sendo negociada a obtenção dos dados de uma vídeo locadora da cidade, mas devido a entraves na política de segurança da empresa que fornece o sistema para esta locadora, não foi possível coletar os dados reais para teste de campo do sistema. A alternativa encaminhada foi utilizar uma base, também real, disponível abertamente na internet. A base de dados utilizada foi a MovieLens Data Set (<http://www.grouplens.org>), fornecida pelo grupo de pesquisa GroupLens Research. Essa base conta com 100.000 avaliações para 1682 filmes por 943 usuários. A base MovieLens foi construída a partir do sítio de recomendações de filmes: movielens.org. A base de dados MovieLens é assim organizada:

- Arquivo u.data: arquivo contendo 100.000 avaliações de 943 usuários para 1682 filmes. Cada usuário avaliou, no mínimo 20 títulos.
- Arquivo u.item: contém informações sobre os filmes, título, data de lançamento e gênero.
- Arquivo u.genre: lista de gêneros de filmes.

- Arquivo u.user: informação demográfica sobre os usuários: nome, idade, gênero, profissão.
- Arquivos de treinamento e teste: a base u.data é dividida em dois tipos de arquivos com a relação de 80%/20% para treinamento e teste, respectivamente.

Para o escopo deste trabalho, não serão utilizadas as avaliações dos usuários durante a fase de treinamento, pois fugiria ao objetivo que é fazer recomendações em ambientes de locadoras de filmes reais, onde não há o conceito de avaliação pelos clientes.

O ambiente de experimentos foi assim determinado:

Da base de filmes foram extraídos o número de identificação único do filme (ID), título do filme, ano de lançamento, gênero e número de vezes que aparece no arquivo de avaliações. Este último parâmetro representa, para a realidade do presente trabalho, o número de locações totais do filme na locadora. O arquivo de avaliações dos usuários representa as locações de cada cliente, portanto o mapa é construído cruzando as informações destas duas tabelas.

O arquivo de testes representa os filmes que o cliente deseja locar, então esses padrões são apresentados ao sistema com a rede já treinada.

4.2 Resultados

Foram criadas tabelas para visualização e estudo dos resultados obtidos com os experimentos realizados.

O primeiro título apresentado nas tabelas é o filme que o cliente deseja locar, os três seguintes são os vizinhos mais próximos no mapa, que indicam ser semelhantes ao primeiro. Para a avaliação dos resultados foi considerada a nota ou avaliação que o usuário deu ao filme. Por exemplo, na Tabela 2, o cliente 1 deseja locar o filme *Shanghai Triad*. Quando apresentado este filme à rede, o cliente obteve como resposta os filmes: *The White Balloon*, *Belle de jour* e *Jean de Florette*, evidenciando factualmente que estes filmes têm similaridades com o primeiro. Observando os títulos que foram mostrados como relacionados, o cliente irá avaliar se vale a pena alugar o filme *Shanghai Triad*. Para avaliar esse resultado, observa-se a coluna **Avaliação** da tabela, os filmes relacionados possuem avaliação de 4, 3 e 5, respectivamente, indicando que o cliente teve satisfação razoável a ótima ao assisti-los. Portanto, a interpretação desse resultado é que o filme

Shanghai Triad provavelmente irá agradar o cliente. A avaliação para o filme *Shanghai Triad* foi de 5, ou seja, a satisfação do usuário foi muito boa.

Tabela 2: Resultados para o cliente 1 e o filme *Shangai Triad*

Título	Gênero	Ano	Número de locações	Avaliação
Shanghai Triad	Drama	1995	20	5
White Balloon, The	Drama	1995	7	4
Belle de jour	Drama	1967	30	3
Jean de Florette	Drama	1986	55	5

O cliente 1 tem 135 títulos em seu histórico, quantidade que rendeu um bom aprendizado para a rede. De forma geral, foi observado que a rede conseguiu gerar um bom perfil para este cliente. A tabela 3 mostra outro resultado para o cliente 1, a análise é feita de modo análogo à tabela anterior. O primeiro título representa o filme que o cliente ainda não locou, os três seguintes são os filmes mais similares ao primeiro. Observa-se que eles possuem alta similaridade em relação ao gênero e ao ano de lançamento. A satisfação do cliente é confirmada pela avaliação que deu aos filmes.

Tabela 3: Resultados para o cliente 1 e o filme *The Usual Suspects*

Título	Gênero	Ano	Número de locações	Avaliação
Usual Suspects, The	Crime/Suspense	1995	211	5
Seven (Se7en)	Crime/Suspense	1995	195	2
Copycat	Crime/Drama/ Suspense	1995	69	3
Four Rooms	Suspense	1995	75	4

A Tabela 4 mostra o resultado para o cliente 2, desejando alugar o filme *Mighty Aphrodite*. Neste exemplo a rede também teve um bom desempenho em relação a encontrar filmes fortemente relacionados e que representam as preferências do cliente. O cliente 2 possui apenas 40 títulos no seu histórico. Observando a coluna Avaliação, que representa a satisfação do usuário, pode-se perceber que o sistema obteve bons resultados para este filme. Considerando esta métrica, os resultados para o cliente 2 foram, na maioria, satisfatórios, tendo o sistema errado poucas vezes no agrupamento dos títulos.

Tabela 4: Resultados para o cliente 2 e o filme *Mighty Aphrodite*

Título	Gênero	Ano	Número de locações	Avaliação
Mighty Aphrodite	Comédia	1995	134	4
Kolya	Comédia	1996	94	5
Birdcage, The	Comédia	1996	231	4
Full Monty, The	Comédia	1997	252	4

Devido à pouca quantidade de padrões para treinamento para o cliente 2, é possível observar na Tabela 5 que o filme *Apt Pupil*, com baixa avaliação, relaciona-se com outros que foram bem aprovados pelo cliente. Obviamente, a nota baixa não necessariamente deve ser encarado como uma evidência de que o sistema errou, pois fatores externos como um filme de tema, atores e produção boa, pode ter sido mal dirigido (na opinião do cliente).

Tabela 5: Resultados para o cliente 2 e o filme *Apt Pupil*

Título	Gênero	Ano	Número de locações	Avaliação
Apt Pupil	Drama	1998	136	1
Wings of the Dove, The	Drama/Romance/ Suspense	1997	65	5
Restoration	Drama	1995	52	4
Promesse, La	Drama	1996	6	3

Nas tabelas seguintes é possível observar os resultados dos experimentos para outros clientes.

Tabela 6: Resultados para o cliente 6 e o filme *Il Postino*

Título	Gênero	Ano	Número de locações	Avaliação
Postino, Il	Drama/Romance	1994	140	5
Chasing Amy	Drama/Romance	1997	203	3
Like Water For Chocolate	Drama/Romance	1992	121	5
Jerry Maguire	Drama/Romance	1996	309	2

Tabela 7: Resultados para o cliente 6 e o filme *Pulp Fiction*

Título	Gênero	Ano	Número de locações	Avaliação
Pulp Fiction	Crime/Drama	1994	312	4
GoodFellas	Crime/Drama	1990	177	4
Donnie Brasco	Crime/Drama	1997	129	3
Godfather, The	Crime/Drama	1972	340	5

Esses resultados experimentais, nos levam a crer que nossa proposta (i.e. o sistema proposto) correspondeu às expectativas, pois conseguiu ser um auxiliar nas decisões do cliente, trazendo boas respostas na maioria dos casos testados.

5 *Conclusões e Trabalhos Futuros*

5.1 Conclusões

O trabalho proposto foi desenvolvido como uma prova de conceito para sistemas de recomendação no ambiente de vídeo locadoras utilizando, para aprendizado do comportamento de clientes, o algoritmo de redes SOM. A contribuição deste trabalho é ajudar o cliente, dando diretrizes para realizar sua escolha.

Sistemas de recomendação são bastante utilizados em sítios de comércio eletrônico e existem diversas ferramentas que recomendam filmes na internet. Os sistemas de informação de vídeo locadoras carecem de ferramentas que auxiliem o cliente quando este tem dúvida de qual filme locar, ou quando simplesmente deseja opiniões de terceiros. Geralmente o cliente busca filmes recomendados por algum amigo ou familiar que já tenha assistido e expressou sua opinião, mas não é simples saber realmente qual tipo de filme irá agradar esse cliente. O modelo apresentado neste trabalho consegue montar um perfil para o usuário com o objetivo de agregar informações às opiniões que ele já obteve. Foi possível construir modelos individualizados e oferecer ao cliente a relação de um novo filme com aqueles que já estão presentes em seu histórico.

5.2 Dificuldades e trabalhos futuros

As redes SOM conseguem trabalhar bem como extrator de características, para isso precisam de várias informações sobre os padrões de entrada. Uma dificuldade no desenvolvimento do trabalho foi obter outros parâmetros associados aos filmes, maiores detalhes que pudessem ser considerados no momento do treinamento da rede. Informações como premiações que o filme recebeu, como Oscar ou Globo de Ouro, quantidade de premiações, ou ainda, diretor, ator principal, entre outros detalhes que não constam na base do MovieLens, mas podem ser facilmente obtidos do banco de dados de uma locadora e poderiam ainda mais melhorar as sugestões. Com estas informações haveria um ajuste

melhor da rede de forma geral, caracterizando melhor o perfil do cliente. A adição dessas características está entre os futuros esforços para melhoria da ferramenta.

Ainda outra melhoria que pode ser incluída no modelo é a criação de um ambiente gráfico mais amigável ao usuário. Uma das características desse ambiente seria um formulário onde o cliente pudesse simplesmente digitar o título do filme a ser locado e o sistema buscar os filmes mais similares a ele. Nesse ambiente o cliente teria também a opção de visualizar o mapa topológico de seu perfil num plano bidimensional, tendo uma visão geral do seu histórico, podendo então, direcionar sua escolha a um título que tenha, por exemplo, o mesmo gênero do agrupamento que possui mais filmes.

Um ponto muito importante que deve ser acrescentado em trabalhos futuros é o armazenamento do mapa de um cliente. Criar mapas para clientes que possuem muitos filmes em seu histórico demanda muito processamento e tempo. Armazenar as informações do mapa do cliente é imprescindível para tornar o processo de recomendação mais eficiente. Neste caso devem ser estudadas técnicas para não precisar refazer todo mapa cada vez que o cliente loca um filme, ou seja, insere um novo título no seu histórico e manter o mapa sempre atualizado.

Referências

- 1 ANDERSON, C. *A Cauda Longa: do mercado de massa para o mercado de nicho*. [S.l.]: Elsevier, 2006.
- 2 FILHO, V. M. e-recommender: Sistema inteligente de recomendação para comércio eletrônico. *Trabalho de Conclusão do Curso de Engenharia da Computação*, Universidade de Pernambuco, 2006.
- 3 TERVEEN, L.; HILL, W. Beyond recommender systems: Helping people help each other. *HCI In The New Millennium*, Addison-Wesley, 2001.
- 4 LYMAN, P.; VARIAN, H. R.; CHARLES, P.; GOOD, N.; JORDAN, L. L.; PAL, J.; SWEARINGEN, K. How much information? 2003. *Technical report*, UC Berkeleys School of Information Management and Systems, 2003. Disponível em: <<http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/>>.
- 5 RESNIK, P.; VARIAN, H. R. Recommender systems. *Communications of the ACM*, v. 40, n. 3, 1997.
- 6 GOLDBERG, D.; NICHOLS, D.; OKI, B. M.; TERRY, D. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, v. 35, n. 12, p. 61–70, 1992.
- 7 BEZERRA, B. L. D. *Uma solução em filtragem de informação para sistemas de recomendação baseada em análise de dados simbólicos*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2004.
- 8 REATEGUI, E. B.; CAZELLA, S. B. Sistemas de recomendação. mini-curso. Enia, 2005.
- 9 BALABANOVIC, M.; SHOHAM, Y. Fab: Content-based, collaborative recommendation. *Communication of the ACM*, v. 40, n. 3, p. 66–72, 1997.
- 10 GABRIELSSON, S.; GABRIELSSON, S. *The use of Self-Organizing Maps in Recommender Systems*. Dissertação (Mestrado) — Uppsala University, 2006.
- 11 HERLOCKER, J. L. *Understanding and Improving Automated Collaborative Filtering Systems*. Tese (Doutorado) — University of Minnesota, 2000.
- 12 DENNING, P. J. Electronic junk. *Communications of the ACM*, v. 25, n. 3, 1982.
- 13 LANG, K. Newsweeder: Learning to filter netnews. In: *12th International Conference on Machine Learning*. [S.l.: s.n.], 1995.

- 14 KRULWICH, B.; BURKEY, C. Learning user information interests through extraction of semantically significant phrases. In: *AAAI Spring Symposium on Machine Learning in Information Access*. [S.l.: s.n.], 1996.
- 15 BURKE, R. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, v. 12, n. 4, p. 331–370, 2002.
- 16 RESNICK, P.; IACOVOU, N.; SUCHAK, M.; BERGSTROM, P.; RIEDL, J. GroupLens: an open architecture for collaborative filtering of netnews. In: *CSCW '94: Proceedings of the 1994 ACM conference on Computer supported cooperative work*. [S.l.: s.n.], 1994. p. 175–186.
- 17 SHARDANAND, U.; MAES, P. Social information filtering: Algorithms for automating "word of mouth". In: *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*. [S.l.: s.n.], 1995. p. 210–217.
- 18 HILL, W.; STEAD, L.; ROSENSTEIN, M.; FURNAS, G. Recommending and evaluating choices in a virtual community of use. In: *CHI '95: Proceedings of the SIGCHI conference on Human factors in computing systems*. [S.l.: s.n.], 1995. p. 194–201.
- 19 CLAYPOOL, M. and GOKHALE, A. and MIRANDA, T. and MURNIKOV, P. and NETES, D. and SARTIN, M. Combining content-based and collaborative filters in an online newspaper. In: *SIGIR '99 Workshop on Recommender Systems - Implementation and Evaluation*. [S.l.: s.n.], 1999.
- 20 SMYTH, B.; COTTER, P. Personalized electronic program guides for digital tv. *AI Magazine*, American Association for Artificial Intelligence, v. 22, n. 2, 2001.
- 21 BASU, C.; HIRSH, H.; W., C. Recommendation as classification: Using social and content-based information in recommendation. In: *Proceedings of the 15th National Conference on Artificial Intelligence*. [S.l.: s.n.], 1998. p. 714–720.
- 22 BRAGA, A. P.; CARVALHO, A. P. L. F.; LUDERMIR, T. B. *Redes neurais artificiais: teoria e aplicações*. [S.l.]: Editora LTC, 2000.
- 23 KOHONEN, T. *Self-Organizing and Associative Memory*. 3. ed. [S.l.]: Springer-Verlag, 1989.
- 24 KOHONEN, T. The self-organizing map. In: *Proceedings of the IEEE*. [S.l.: s.n.], 1990. v. 78, n. 9, p. 1464–1480.
- 25 KOHONEN, T. The "neural" phonetic typewriter. *IEEE Computer*, v. 21, n. 3, p. 11–22, 1988.
- 26 KOHONEN, T. and KASKI, S. and LAGUS, K. and SALOJÄRVI, J. and HONKELA, J. and PAATERO, V. and SAARELA, A. Self organizing of a massive document collection. *IEEE Transactions on Neural Networks*, v. 11, n. 3, 2000.
- 27 GRAEF, G.; SCHAEFER, C. Application of art2 networks and self-organizing maps to collaborative filtering. *Revised Papers from the International Workshops OHS-7, SC-3, and AH-3 on Hypermedia: Openness, Structural Awareness, and Adaptivity*, p. 182–184, 2002.

- 28 LEE, M.; CHOI, P.; WOO, Y. A hybrid recommender system combining collaborative filtering with neural network. In: *AH '02: Proceedings of the Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*. [S.l.: s.n.], 2002. p. 531–534.

Apêndice A: Tabelas de resultados

Como complemento do Capítulo 4, este apêndice apresenta tabelas com resultados para os demais experimentos.

Tabela 8: Resultados para o cliente 1 e o filme *Richard III*

Título	Gênero	Ano	Número de locações	Avaliação
Richard III	Drama/Guerra	1995	73	3
Bridge on the River Kwai, The	Drama/Guerra	1957	130	4
Gone with the Wind	Drama/Romance/ Guerra	1939	138	4
Lone Star	Drama/Mistério	1996	148	5

Tabela 9: Resultados para o cliente 1 e o filme *Postino, Il*

Título	Gênero	Ano	Número de locações	Avaliação
Postino, Il	Drama/Romance	1994	140	5
Breakfast at Tiffany's	Drama/Romance	1961	77	1
Jerry Maguire	Drama/Romance	1996	309	2
Graduate, The	Drama/Romance	1967	192	5

Tabela 10: Resultados para o cliente 1 e o filme *Rock, The*

Título	Gênero	Ano	Número de locações	Avaliação
Rock, The	Ação/Aventura/ Suspense	1996	302	3
GoldenEye	Ação/Aventura/ Suspense	1995	105	3
Muppet Treasure Island	Ação/Aventura/ Comédia/Musical/ Suspense	1996	68	1
Heavy Metal	Ação/Aventura/ Animação/Terror/ Ficção Científica	1967	54	5

Tabela 11: Resultados para o cliente 1 e o filme *Platoon*

Título	Gênero	Ano	Número de locações	Avaliação
Platoon	Drama/Guerra	1986	99	4
Bridge on the River Kwai, The	Drama/Guerra	1957	130	4
Gone with the Wind	Drama/Guerra/Romance	1939	138	4
Lone Star	Drama/Mistério	1996	148	5

Tabela 12: Resultados para o cliente 1 e o filme *Platoon*

Título	Gênero	Ano	Número de locações	Avaliação
Platoon	Drama/Guerra	1986	99	4
Bridge on the River Kwai, The	Drama/Guerra	1957	130	4
Gone with the Wind	Drama/Guerra/Romance	1939	138	4
Lone Star	Drama/Mistério	1996	148	5

Tabela 13: Resultados para o cliente 2 e o filme *Antonia's Line*

Título	Gênero	Ano	Número de locações	Avaliação
Antonia's Line	Drama	1995	59	3
Marvin's Room	Drama	1996	61	3
Promesse, La	Drama	1996	6	3
Good Will Hunting	Drama	1997	154	5

Tabela 14: Resultados para o cliente 2 e o filme *The River Wild*

Título	Gênero	Ano	Número de locações	Avaliação
River Wild, The	Ação/Suspense	1994	113	3
Breakdown	Ação/Suspense	1997	60	4
Heat	Ação/Crime/ Suspense	1995	185	4
Scream	Terror/Suspense	1996	391	3

Tabela 15: Resultados para o cliente 6 e o filme *Apollo 13*

Título	Gênero	Ano	Número de locações	Avaliação
Apollo 13	Ação/Drama/ Suspense	1995	222	2
Fugitive, The	Ação/Suspense	1993	267	3
Face/Off	Ação/Suspense/ Ficção Científica	1997	152	3
Usual Suspects, The	Crime/Suspense	1995	211	4

Tabela 16: Resultados para o cliente 6 e o filme *Forrest Gump*

Título	Gênero	Ano	Número de locações	Avaliação
Forrest Gump	Comédia/Romance/ Guerra	1994	259	3
Truth About Cats & Dogs, The	Comédia/Romance	1996	234	2
Sabrina	Comédia/Romance	1995	152	4
Cinema Para- diso	Comédia/Drama/ Romance	1988	104	4

Tabela 17: Resultados para o cliente 6 e o filme *The Remains of the Day*

Título	Gênero	Ano	Número de locações	Avaliação
Remains of the Day, The	Drama	1993	119	3
Rainmaker, The	Drama	1997	124	2
12 Angry Men	Drama	1957	102	4
Raging Bull	Drama	1980	98	4

Tabela 18: Resultados para o cliente 6 e o filme *Aladdin*

Título	Gênero	Ano	Número de locações	Avaliação
Aladdin	Animação/Infantil/ Comédia/Musical	1992	184	2
Lion King, The	Animação/Infantil/ Musical	1994	175	4
Mary Poppins	Infantil/Comédia/ Musical	1964	147	4
Toy Story	Animação/Infantil/ Comédia	1995	383	4

Tabela 19: Resultados para o cliente 11 e o filme *Dead Man Walking*

Título	Gênero	Ano	Número de locações	Avaliação
Dead Man Wal- king	Drama	1995	246	5
Mr. Holland's Opus	Drama	1995	235	5
Dead Poets Soci- ety	Drama	1989	197	5
People vs. Larry Flynt, The	Drama	1996	166	4

Tabela 20: Resultados para o cliente 11 e o filme *Braveheart*

Título	Gênero	Ano	Número de locações	Avaliação
Braveheart	Drama/Guerra	1995	230	4
Schindler's List	Drama/Guerra	1993	237	5
Apocalypse Now	Drama/Guerra	1979	179	2
English Patient, The	Drama/Romance/Guerra	1996	388	5

Tabela 21: Resultados para o cliente 11 e o filme *Bob Roberts*

Título	Gênero	Ano	Número de locações	Avaliação
Bob Roberts	Comédia	1992	65	4
Brady Bunch Movie, The	Comédia	1995	60	3
Fierce Creatures	Comédia	1997	76	3
Jeffrey	Comédia	1995	27	4

Tabela 22: Resultados para o cliente 11 e o filme *2001: A Space Odyssey*

Título	Gênero	Ano	Número de locações	Avaliação
2001: A Space Odyssey	Drama/Mistério/Ficção Científica/Suspense	1968	202	4
Client, The	Drama/Mistério/Suspense	1994	82	4
Rear Window	Mistério/Suspense	1954	173	4
Remains of the Day, The	Drama	1993	119	4